

Cost Effective Rumor Containment in Social Networks

Bhushan Kotnis and Joy Kuri

Dept. of Electronic Systems Engineering, Indian Institute of Science, Bangalore. {bkotnis,kuri}@dese.iisc.ernet.in

Abstract—The spread of rumors through social media and online social networks can not only disrupt the daily lives of citizens but also result in loss of life and property. A rumor spreads when individuals, who are unable to decide the authenticity of the information, mistake the rumor as genuine information and pass it on to their acquaintances. We propose a solution where a set of individuals (based on their degree) in the social network are trained and provided resources to help them distinguish a rumor from genuine information. By formulating an optimization problem we calculate the optimum set of individuals, who must undergo training, and the quality of training that minimizes the expected training cost and ensures an upper bound on the size of the rumor outbreak. Our primary contribution is that although the optimization problem turns out to be non convex, we show that the problem is equivalent to solving a set of linear programs. This result also allows us to solve the problem of minimizing the size of rumor outbreak for a given cost budget. The optimum solution displays an interesting pattern which can be implemented as a heuristic. These results can prove to be very useful for social planners and law enforcement agencies for preventing dangerous rumors and misinformation epidemics.

I. INTRODUCTION

The past decade has seen a dramatic increase in the usage of online social networking and microblogging services [1] like Facebook, Google+, Twitter, etc. Apart from their usefulness in helping individuals keep in touch, they are increasingly being used for disseminating information about events happening in real time [2]. Due to the proliferation of smart phones, individuals can easily capture the unfolding of events in real time and can share it with others via social media instantaneously. Social media played a key role in sharing of information in the immediate aftermath of: the Fukushima nuclear accident [3], hurricane Sandy [4] and the Boston marathon bombings [5]. Online social networks, microblogging, and short messaging services can be extremely useful in aiding the dissemination of time critical and potentially life saving information during large scale human emergencies.

However, along with useful information, online social networks can also aid the spread of rumors, and can even sometimes start a potentially dangerous misinformation epidemic [4]. Since these services operate in a decentralized manner, i.e., absence of a central authority for guaranteeing the authenticity of the information, careless individuals can accidentally initiate and propagate rumors. Furthermore, the distributed nature of social media can be exploited by malicious agents for spreading dangerous misinformation epidemics. This problem has been a growing concern for administrative authorities. In the recent past, with the goal of preventing the spread of

rumors, the Indian Government ordered cell phone operators to impose a limit on the number of text messages that can be sent by any individual [6]. However, such ad-hoc measures are costly and also not very effective. In this article we propose a cost effective rumor prevention mechanism which provides guarantees on the size of the rumor outbreak.

The problem of maximizing the spread of information in social networks is well known [7–9]. Approaches as diverse as algorithmic complexity [7, 8] and optimal control [10, 11] have been proposed. The reverse problem of limiting the spread of rumors has also received some attention. One of the well studied approach is to identify and recruit a set of individuals who will spread an ‘anti-rumor’ [12–14] to combat the rumor. Other approaches include modeling the rumor as a Susceptible Infected Recovered (SIR) or Susceptible Infected Susceptible (SIS) epidemic and using vaccination strategies for limiting the spread [15, 16]. The problem of limiting the spread of virus or malware in computer networks is analogous to the problem of limiting the spread of rumors, and studies [17, 18] have used optimal control techniques for achieving the same. As the saying goes “there is no such thing as a free lunch”, any kind of intervention comes with a cost and most approaches other than the optimal control approach do not consider the economic cost of preventing the spread of rumor. Furthermore, the optimal control solution is difficult to implement as it requires a centralized real time controller.

Here, we propose a method in which a set of individuals are recruited and trained to distinguish rumors from useful information. Such training may also involve allocating costly resources, such as access to real time satellite or drone photography, which help them make the correct decision. Thus, recruiting and training individuals is a costly affair. We seek to identify the set of individuals to be trained, that minimizes the expected training cost and prevents a rumor outbreak by formulating and solving an optimization problem. We also address the problem of minimizing the size of the rumor outbreak for a given cost budget. This is different from the algorithmic approach [13] of finding the optimal set of nodes for starting the ‘anti-rumor’ [12]. Our approach is similar to the one which involves vaccinating people against rumors [15], but there are a few key differences. The cost of vaccination is same for all individuals, while the cost of recruiting and training individuals may not be homogeneous across all individuals. Although training can help increase the accuracy of guessing the nature of the message, it cannot drive down the error probability to near zero levels. This is very different from vaccines, as they have a very high efficacy.

Our model assumes two types of individuals viz. trained and ignorant, and they are connected to one another through a social network. Both, the ignorant and the trained are unable to distinguish between the rumor with certainty, however trained individuals can distinguish a rumor from an information better than the untrained ones. Individuals are characterized by their degree (number of connections) and the cost of training is assumed to be proportional to their degree. Given the set of individuals to be trained, we first calculate the size of the rumor outbreak using a branching process approximation, and then find the set which minimizes the cost for a given outbreak size and the minimum outbreak size for given cost budget. The optimization problem is found to be a non linear program. Our primary contribution is, that we show that the nonlinear problem can be addressed by solving one or more linear programs. Furthermore, we discover that the set of individuals which need to be trained displays an interesting pattern: it turns out that low degree individuals are more important than high degree individuals for the purpose of rumor prevention.

Our contributions are summarized as follows :

- We calculate the size of a rumor outbreak using a branching process approximation.
- We formulate an optimization problem and show that the nonlinear optimization problem can be solved by solving one or more linear programs.
- The solution of the optimization problem displays interesting patterns which can be converted into an implementable heuristic.

II. MODEL

We divide the total population of N individuals into two types: the ignorant (type 1) and the trained (type 2). These individuals are connected with one another through a social network, which is represented by a graph (network). Nodes represent individuals while a link embodies the communication pathways between individuals. For the sake of analytical tractability we make the following approximation. Instead of analyzing the adjacency matrix of the network, we obtain statistical information about the social network by calculating its degree distribution (probability that a randomly chosen node has k neighbors). We then generate a synthetic network with the obtained degree distribution using the configuration model procedure [19]. A sequence of N integers, called the degree sequence, is obtained by sampling the degree distribution. Thus each node is associated with an integer which is assumed to be the number of half edges or stubs associated with the node. Assuming that the total number of stubs is even, each stub is chosen at random and joined with another randomly selected stub. This process continues until all stubs are exhausted. Self loops and multiple edges are possible, but the number of such self loops and multiple edges goes to zero as $N \rightarrow \infty$ with high probability. The network obtained by this process is termed as configuration model.

Let $P(k)$ be the degree distribution of the social network. Individuals, both trained and ignorant can receive a messages which may be benign or malicious (rumors). We assume that the rumor deals with a single and specific topic, and hence

we represent it as a data message, i.e., rumor message R . We propose an approach where the social planners recruit and train individuals with the goal helping them recognize rumors from true information. Let $\phi(k)$ be the proportion of individuals with k degrees recruited for training.

We assume that initially a randomly chosen individual acts as a seed and transmits the R message to *all* its neighbors with probability 1. Thus, only the rumor message is circulating in the social network. The individual who receives this message cannot recognize with certainty if it is a rumor. Each individual makes a hypothesis about the message. An individual decides to transmit a message to all her neighbors if she believes that it is not a rumor. If she thinks the message is a rumor then she does not transmit it to any of her neighbors. An individual who receives the message and decides to spread it to her neighbors, does so only once and such a person is termed as a *believer*. This models the scenario where an individual may hear about an event and report it to all her friends based on her gut feeling about the nature of the message. A believer cannot revert back to being an ignorant, while an ignorant who receives a message and concludes its a rumor is assumed to remain ignorant.

Let H_0 be the hypothesis that the received message is not R , while H_1 be the hypothesis that received message is R . Thus an individual who receives a rumor message transmits it with the probability $\mathbb{P}\{H_0 | \text{it is a rumor message}\}$, i.e., it is the probability of the event that she fails to identify the true nature of the message. Let T_1 and T_2 be the probability of misclassifying the rumor as information for ignorant and trained individuals respectively. We refer to T_1 as the force of rumor. Due to the training $T_2 < T_1$. However, no amount of training can accurately identify the nature of the message, hence $T_1, T_2 \in (0, 1)$.

We present the following scenario as an illustrative example. An ignorant individual A receives a rumor message, she makes an error in identifying it as a rumor, and hence passes on this message to all her connections. This event happens with probability T_1 . A trained Individual B , receives this message from A , but unlike A she correctly identifies the message as a rumor and hence does not send it to her connections. This event happens with probability $1 - T_2$. The proportion of believers, after the process terminates, is termed as size of the outbreak.

III. ANALYSIS

We approximate the rumor spreading process as a branching process [20]. Let $P(k' | k)$ be the probability of encountering a node of degree k' by traversing a randomly chosen link from a node of degree k . In other words, $P(k' | k)$ is the probability that a node with degree k has a neighbor with degree k' . For a network generated by configuration model, $P(k' | k) = \frac{k' P(k')}{\langle k \rangle}$ [21], which is independent of k , where $\langle k^i \rangle$ is the i^{th} moment of $P(k)$.

Let q be the probability of encountering a trained node by traversing a randomly chosen link from a node of degree k . Therefore, $q = \sum_{k'=1}^{\infty} Pr(\text{Neighboring node is trained} | \text{neighboring node has degree } k') \cdot Pr(\text{Neighboring node has$

degree $k' \mid$ original node has degree k).

$$q = \frac{1}{\langle k \rangle} \sum_{k=1}^{\infty} k \phi(k) P(k)$$

The probability that a randomly chosen node has k_1 ignorant and k_2 trained neighbors $= P(k_1, k_2) = \sum_{k: k=k_1+k_2} Pr(k_1, k_2 \mid \text{node has degree } k) P(k)$.

Consider two nodes, A and B , having a common neighbor C . Let X and Y be the number of degrees of A and B respectively. If there is no link connecting A to B , then random variables X and Y are independent. This is because the configuration model is constructed by generating a degree sequence and the N samples which generate the degree sequence are drawn independently. If there is a link connecting A and B (A, B, C forms a triangle), then $P(Y = y \mid X = x) = \frac{yP(y)}{\sum_y yP(y)}$. Thus, X does not provide any knowledge about Y and vice versa. Furthermore, the number of triangles in a network, generated by the configuration model, decays to 0 as $N \rightarrow \infty$ [21]. This suggests that X is independent of Y . The probability that a node is trained, is a function of its degree, hence the event that A is trained (ignorant) is independent of the event that B is trained (ignorant). This allows us to write :

$$P(k_1, k_2) = \binom{k_1 + k_2}{k_2} q^{k_2} (1 - q)^{k_1} P(k_1 + k_2)$$

Due to this independence, $P(k_1, k_2) = Pr(\text{node has } k_1 \text{ ignorant neighbors}) \cdot Pr(\text{node has } k_2 \text{ trained neighbors}) = P_1(k_1)P_2(k_2)$.

Let $Q(k)$ be the excess degree distribution, i.e., the degree distribution of a node arrived at by following a randomly chosen link without counting that link. For the configuration model $Q(k) = \frac{(k+1)P(k+1)}{\langle k \rangle}$. Let $Q(k_1, k_2)$ be the excess degree distribution for connections to ignorant and trained nodes.

$$Q(k_1, k_2) = \binom{k_1 + k_2}{k_2} q^{k_2} (1 - q)^{k_1} Q(k_1 + k_2)$$

Also $Q(k_1, k_2) = Q_1(k_1)Q_2(k_2)$, where $Q_1(k_1)$ and $Q_2(k_2)$ are the excess degree distribution counterparts of $P_1(k_1)$ and $P_2(k_2)$ respectively.

Let $\tilde{P}(k)$ and $\tilde{Q}(k)$ be the distribution and the excess distribution of the number of neighbors who believe in the rumor. Similarly let $\tilde{P}(\tilde{k}_1, \tilde{k}_2)$ and $\tilde{Q}(\tilde{k}_1, \tilde{k}_2)$ be the distribution and the excess distribution of the number of ignorant and trained neighbors who are believers. An ignorant becomes a believer with probability T_1 while a trained node becomes a believer with probability T_2 .

$$\begin{aligned} \tilde{P}(\tilde{k}_1, \tilde{k}_2) &= \sum_{k_1=\tilde{k}_1}^{\infty} \sum_{k_2=\tilde{k}_2}^{\infty} P(k_1, k_2) \prod_{i=1}^2 \binom{k_i}{\tilde{k}_i} T_i^{\tilde{k}_i} (1 - T_i)^{k_i - \tilde{k}_i} \\ \tilde{Q}(\tilde{k}_1, \tilde{k}_2) &= \sum_{k_1=\tilde{k}_1}^{\infty} \sum_{k_2=\tilde{k}_2}^{\infty} Q(k_1, k_2) \prod_{i=1}^2 \binom{k_i}{\tilde{k}_i} T_i^{\tilde{k}_i} (1 - T_i)^{k_i - \tilde{k}_i} \end{aligned}$$

Generating function	Distribution
$G(u_1, u_2)$	$P(k_1, k_2)$
$F(u_1, u_2)$	$Q(k_1, k_2)$
$\tilde{G}(u_1, u_2)$	$\tilde{P}(\tilde{k}_1, \tilde{k}_2)$
$\tilde{F}(u_1, u_2)$	$\tilde{Q}(\tilde{k}_1, \tilde{k}_2)$
$G_i(u)$	$P_i(k)$
$\tilde{G}_i(u)$	$\tilde{P}_i(\tilde{k})$
$F_i(u)$	$Q_i(k)$
$\tilde{F}_i(u)$	$\tilde{Q}_i(\tilde{k})$
$G_0(u)$	$P(k)$
$F_0(u)$	$Q(k)$
$\tilde{G}_0(u)$	$\tilde{P}(\tilde{k})$
$\tilde{F}_0(u)$	$\tilde{Q}(\tilde{k})$

TABLE I
LIST OF PROBABILITY GENERATING FUNCTIONS.

Also, $\tilde{P}(\tilde{k}_1, \tilde{k}_2) = \tilde{P}_1(\tilde{k}_1)\tilde{P}_2(\tilde{k}_2)$ and $\tilde{Q}(\tilde{k}_1, \tilde{k}_2) = \tilde{Q}_1(\tilde{k}_1)\tilde{Q}_2(\tilde{k}_2)$. Where $\tilde{P}_i(\tilde{k}_i)$ and $\tilde{Q}_i(\tilde{k}_i)$ are the marginals. The probability generating functions, $G(u) = \sum_{k=0}^{\infty} u^k P(k)$, for the distributions discussed above are listed in Table I. Now, $\tilde{G}(u_1, u_2)$ is given by

$$\begin{aligned} &\sum_{\tilde{k}_1, \tilde{k}_2} u_1^{\tilde{k}_1} u_2^{\tilde{k}_2} \sum_{k_1=\tilde{k}_1}^{\infty} \sum_{k_2=\tilde{k}_2}^{\infty} P(k_1, k_2) \prod_{i=1}^2 \binom{k_i}{\tilde{k}_i} T_i^{\tilde{k}_i} (1 - T_i)^{k_i - \tilde{k}_i} \\ &= \sum_{k_1, k_2} (1 + (u_1 - 1)T_1)^{k_1} (1 + (u_2 - 1)T_2)^{k_2} P(k_1, k_2) \\ &= G(1 + (u_1 - 1)T_1, 1 + (u_2 - 1)T_2) \end{aligned}$$

Similarly, $\tilde{F}(u_1, u_2) = F(1 + (u_1 - 1)T_1, 1 + (u_2 - 1)T_2)$ and $\tilde{G}_i(u) = G_i(1 + (u - 1)T_i)$. Also note that $\tilde{G}_0(u) = \tilde{G}_1(u)\tilde{G}_2(u)$. This is because the total number of believing neighbors is the sum of trained and untrained neighbors who are believers, and the fact that $\tilde{P}_1(k_1)$ is independent of $\tilde{P}_2(k_2)$.

In a two phase branching process, the distribution of the number of children in the first generation is different than the distribution of children in subsequent generations. In our case the children are equivalent to believers. At time $t = 0$, the seed spreads rumor to all her neighbors, however all her neighbors may not become believers. The probability that \tilde{k} connections become believers is given by $\tilde{P}(\tilde{k})$. For the subsequent generations the probability of \tilde{k} neighbors is $\tilde{Q}(\tilde{k})$, since $\tilde{Q}(\tilde{k})$ is the distribution on the number of believers encountered after following a link (without counting the link).

Let $\tilde{\mu}$ be the mean number of believers in the first generation and let $\tilde{\nu}$ be the mean number of believers in subsequent generations.

$$\begin{aligned} \tilde{\mu} &= \left. \frac{d}{du} \tilde{G}_0(u) \right|_{u=1} = \left. \frac{d}{du} \tilde{G}_1(u)\tilde{G}_2(u) \right|_{u=1} = T_1 G'_1(1) + T_2 G'_2(1) \\ &= T_1 \sum_{k_1, k_2} k_1 P(k_1, k_2) + T_2 \sum_{k_1, k_2} k_2 P(k_1, k_2) \end{aligned}$$

Similarly,

$$\tilde{\nu} = T_1 \sum_{k_1, k_2} k_1 Q(k_1, k_2) + T_2 \sum_{k_1, k_2} k_2 Q(k_1, k_2)$$

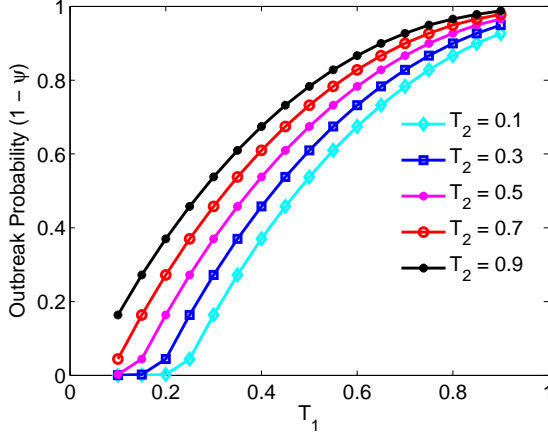


Fig. 1. Outbreak probability $1 - \psi$ vs. T_1 for various values of T_2

From Theorem 3.1.3 in [22] a giant component exists if $\tilde{\nu} > 1$ and its size is asymptotically given by $1 - \tilde{G}_0(u^*)$ where u^* is the smallest fixed point of $u = \tilde{F}_0(u)$. In other words an outbreak is possible only when $\tilde{\nu} > 1$ and the size of such an outbreak is given by $1 - \psi$ where

$$\begin{aligned} \psi &= \tilde{G}_0(u^*) = \tilde{G}_1(u^*)\tilde{G}_2(u^*) \\ &= G_1(1 + (u^* - 1)T_1)G_2(1 + (u^* - 1)T_2) \\ &= \sum_{k_1, k_2}^{\infty} (1 + (u^* - 1)T_1)^{k_1} (1 + (u^* - 1)T_2)^{k_2} P(k_1, k_2) \end{aligned}$$

where u^* is a solution of

$$u = \sum_{k_1, k_2}^{\infty} (1 + (u - 1)T_1)^{k_1} (1 + (u - 1)T_2)^{k_2} Q(k_1, k_2)$$

Clearly, $u^* = 1$ is a solution to the above fixed point equation, but it may not be the smallest fixed point. If $\tilde{\nu} < 1$ then $\psi = 1$ and $u^* = 1$. However if $u^* < 1$ is a fixed point then $\psi < 1$.

We have obtained the expression for the size of the outbreak which can now be used for formulating the optimization problem. The size of the outbreak can also be interpreted as the probability that a randomly chosen individual is a believer, or as the probability of a rumor outbreak [23].

IV. RESULTS

Recruiting, training and equipping individuals with resources to distinguish between rumor and benign message is costly. A high degree individual is more likely to receive a message than a low degree individual. For a trained individual, reception of a message translates into usage of costly resources for making a decision. Hence, we assume that the cost incurred in training an individual is an increasing function of its degree k . Increasing the quality of training, or equivalently, decreasing T_2 , results in decrease of the outbreak probability $1 - \psi$. This is shown graphically in Fig. 1 and analytically in Lemma A.4 detailed in Appendix. Since, increasing the quality of training results in a higher cost, the cost function must be a decreasing function of T_2 . Let $c(k, T_2)$ be the cost of training a node with degree k . The average cost is given

by $\sum_{k=1}^{\infty} c(k, T_2) Pr(\text{node is selected for training} \mid \text{node has degree } k) P(k) = \sum_{k=1}^{\infty} c(k, T_2) \phi(k) P(k)$. The average number of trained individuals is given by $\sum_{k=1}^{\infty} \phi(k) P(k)$.

We formulate two optimization problems, viz., one which minimizes cost while enforcing an upper bound on the outbreak probability, and the other which minimizes the outbreak probability for a given cost budget.

A. Cost minimization problem

Providing guarantees on rumor outbreak at a minimum cost is appropriate in scenarios where the rumor spread may result in loss of life and property, such as rumors that incite communal violence. The guarantee on rumor outbreak probability is written as a constraint to the optimization problem. The cost $c(\phi, T_2)$ is minimized subject to $1 - \psi \leq \delta$ where $\gamma \in [0, 1]$. If $\gamma = 0$, the constraint becomes $\tilde{\nu} \leq 1$, as $\gamma = 0$ implies $\psi = 1$ which is the same as $\tilde{\nu} \leq 1$. For a fixed T_2 , the constraint $1 - \psi \leq \delta$ is non linear in ϕ . For a fixed T_2 , the following theorem allows us to write the non linear constraint as a linear constraint. The proof follows from Lemmas A.1, A.2 and A.3 detailed in the Appendix.

Theorem IV.1. *If $T_2 < T_1$, for $\psi \in [0, 1]$, then ψ is strictly increasing with respect to q , i.e., $\frac{d\psi}{dq} > 0$ for all $q \in [0, 1]$ and $\tilde{\nu}$ is strictly decreasing with respect to q , i.e., $\frac{d\tilde{\nu}}{dq} < 0$, $\forall q \in [0, 1]$, where $q = \frac{1}{\langle k \rangle} \sum_{k=1}^{\infty} k \phi(k) P(k)$.*

Since, $\frac{d\psi}{dq} > 0$, the outbreak probability constraint can be written as $\frac{1}{\langle k \rangle} \sum_{k=1}^{\infty} k \phi(k) P(k) \geq q^*$, where $\psi(q) \big|_{q=q^*} = 1 - \gamma$.

The optimization problem is described by:

$$\begin{aligned} &\text{minimize}_{\phi, T_2} \quad \sum_{k=1}^{\infty} c(k, T_2) \phi(k) P(k) \\ &\text{subject to:} \\ &\quad \frac{1}{\langle k \rangle} \sum_{k=1}^{\infty} k \phi(k) P(k) \geq q^* \\ &\quad \sum_{k=1}^{\infty} \phi(k) P(k) \leq B \\ &\quad T_L \leq T_2 \leq T_u \\ &\quad 0 \leq \phi \leq 1 \end{aligned} \tag{1}$$

B is the upper bound on the average number of individuals that can be trained, $B \in [0, 1]$. $T_l > 0$ is the upper bound on the quality of training and $T_u < T_1$ is the lower bound. The above problem is non linear in T_2 as q^* is a non linear function of T_2 , however, for a fixed T_2 it is a linear program. We convert the problem to a set of linear programs by discretizing T_2 and formulating a linear program for each value of T_2 . We obtain an approximate global minimum by comparing the minimas obtained by solving the set of linear programs.

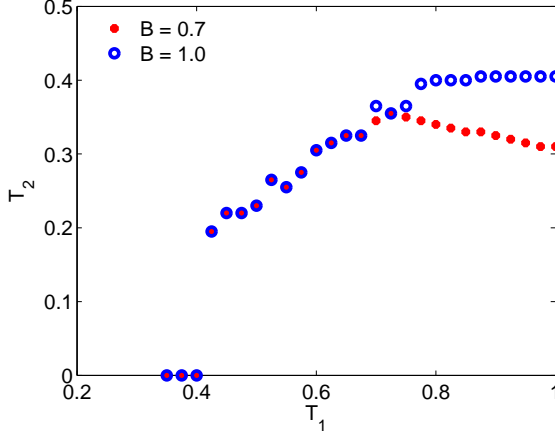


Fig. 2. Optimum value of T_2 vs. T_1 . Parameters: $\gamma = 0.1$, $T_l = 0$, $T_u = T_1$

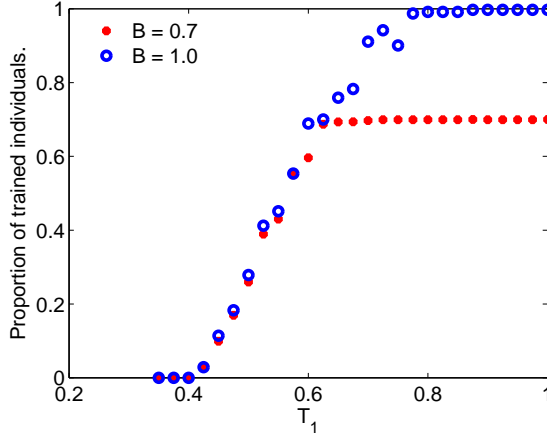


Fig. 3. Proportion of Trained Individuals vs. T_1 . Parameters: $\gamma = 0.1$, $T_l = 0$, $T_u = T_1$

The optimization problem described above may not be feasible for all values of T_1 and for all possible degree distributions $P(k)$. Here is an example of a scenario where the constraint set is empty. Assume $B = 1$, the problem becomes infeasible when $1 - \psi \geq \gamma$ when T_2 is at the minimum possible value (T_l) and $q = 1$, i.e., all individuals are trained. However, if the constraint on the quality of training is relaxed, $T_l = 0$, then an optimal feasible solution will always exist as T_2 can be pushed arbitrarily close to 0 to ensure that the outbreak probability constraint is not violated.

Assuming feasibility of the problem, we use a numerical linear programming solver to arrive at the solution. Since many social networks are scale free [1, 24], we assume that the network degree distribution is a power law $P(k) \propto k^{-\alpha}$. In the numerical results we have assumed $\alpha = 2.5$, a population size of 2000 and cost function which is linear in k , $c(k, T_2) = \frac{k}{T_2}$.

The optimum value of T_2 for varying T_1 is shown in Fig. 2, while Fig. 3 shows the optimum proportion of trained individuals for varying T_1 . For the $B = 1$ scenario, as the force of rumor T_1 increases, the optimum T_2 rises and saturates. Notice that when T_2 saturates the proportion of trained hits 1, i.e., all nodes are trained. Since all the nodes are trained

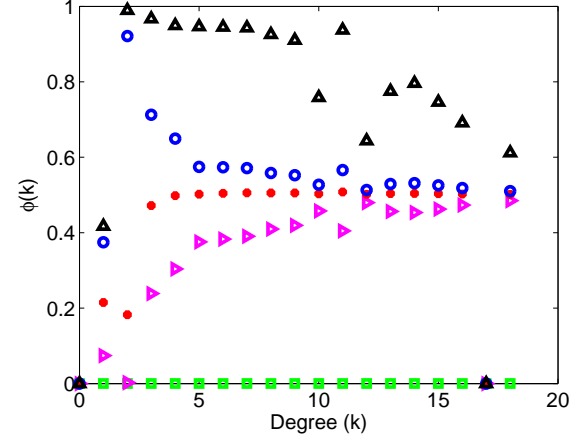


Fig. 4. ϕ for various values of T_1 . Parameters : $\gamma = 0.1$, $B = 0.7$, $T_l = 0$, $T_u = T_1$. Green squares : $T_1 = 0.4$, magenta triangles : $T_1 = 0.45$, red stars : $T_1 = 0.5$, blue circles : $T_1 = 0.6$, black hats : $T_1 = 0.7$.

increase in T_1 does not have any effect on T_2 . This happens because the cost of increasing quality (lower T_2) is higher than the cost of training individuals. However, before the saturation point there is some trade off between the cost of training more nodes and the cost of increasing the quality of training.

When $B = 0.7$, the proportion of trained nodes cannot exceed 0.7 and hence after the proportion of individuals hit 0.7 the only choice to combat the increase in T_1 is increasing the quality of training (reducing T_2). For a lower values of T_1 (before the saturation point) an increase in T_2 is possible as more nodes can be trained, but once the proportion of trained individuals hits a the limit governed by B , T_2 decreases. When T_1 is small (≤ 0.4), $\phi = \emptyset$, thus the cost is 0, and hence T_2 can take any arbitrary value.

Fig. 4 shows ϕ for various values of T_1 . As T_1 increases the proportion of trained low degree nodes increases much faster than the proportion of trained high degree nodes. A clear pattern is seen, the proportion of high degree nodes that need to be trained is more or less constant with respect to T_1 , while the trained low degree nodes are extremely sensitive to T_1 . Thus, one can formulate a simple policy of training a fixed proportion (say 40–60%) of high degree nodes, and recruit low degree nodes depending on the estimated severity of the rumor T_1 . In other words, after fixing the proportion of high degree individuals that are trained, if the social planners perceive that a particular rumor message can be easily identified by individuals then they need not train a whole lot of low degree nodes. Thus, as far as rumor prevention is concerned, low degree nodes are more important than high degree nodes

B. Outbreak probability minimization problem

We now look at the problem of minimizing the outbreak probability in a resource constrained scenario. More, specifically we study the situation where along with a limited cost budget the social planner has very little freedom on modulating the quality of training, i.e., T_2 is fixed. Thus the outbreak probability $1 - \psi$ must be minimized subject to a

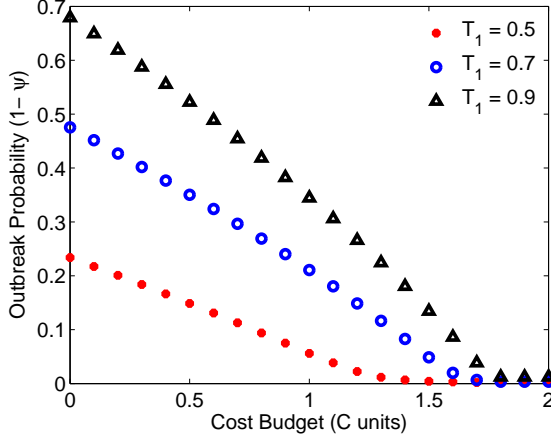


Fig. 5. Outbreak Probability vs. Cost Budget C for various T_1 . Parameters: $T_2 = 0.25$, $B = 0.7$

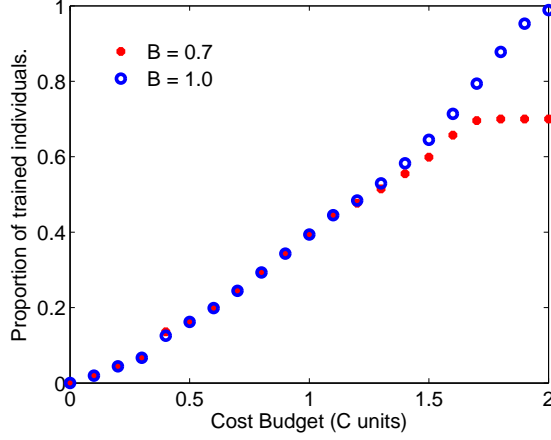


Fig. 6. Proportion of Trained Individuals vs. Cost Budget C .

cost constraint. Since $\frac{d\psi}{dq} > 0$, maximizing q is equivalent to minimizing $1 - \psi$. The optimization problem can be stated as follows:

$$\begin{aligned}
 & \underset{\phi}{\text{maximize}} && \sum_{k=1}^{\infty} k\phi(k)P(k) \\
 & \text{subject to:} && \\
 & \sum_{k=1}^{\infty} c(k)\phi(k)P(k) \leq C && (2) \\
 & \sum_{k=1}^{\infty} \phi(k)P(k) \leq B \\
 & 0 \leq \phi \leq 1
 \end{aligned}$$

We assume that the network degree distribution is a power law $P(k) \propto k^{-\alpha}$. In the numerical results we have assumed $\alpha = 2.5$, total population size of 2000 and cost function which is linear in node degree, i.e., $c(k) = k$. Fig. 5 illustrates the reduction in the outbreak probability with increase in cost budget C . Fig. 6 shows the expected number of trained individuals for varying cost budget. In Fig. 7 we plot ϕ for various values of C . Similar to the previous optimization

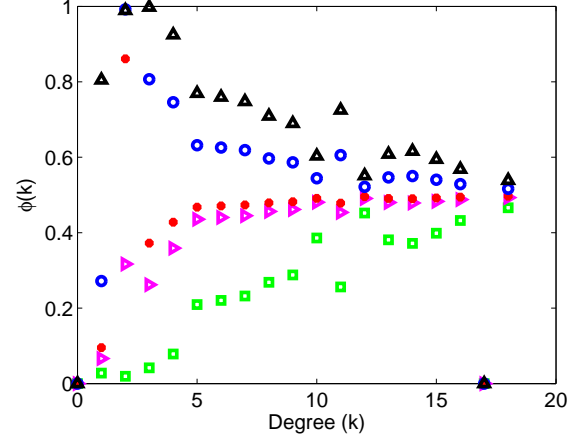


Fig. 7. ϕ for various values of cost budget C . Parameters: $B = 0.7$. Green squares: $C = 0.2$, magenta triangles: $C = 0.6$, red stars: $C = 1.0$, blue circles: $C = 1.4$, black hats: $C = 1.8$.

problem, low degree nodes are seen to be more sensitive to changes in C than high degree nodes.

C. A remark on the information spreading problem

The analysis and results discussed in this article can also be applied to address the problem of spreading information in a social network. Social planners, health campaigners, or political parties may want to ensure the dissemination of information in the social network at minimal cost. Incentivizing individuals for spreading information can help in disseminating the message to a large number of people. Redefine T_1 as the probability an ordinary individual shares the message, and T_2 as the probability that an recruited individual shares the message ($T_2 > T_1$). Let $\phi(k)$ be the proportion of recruited individuals with degree k , and $c(k, T_2)$ be the recruitment cost.

An optimization problem for minimizing the cost guaranteeing the size of the information outbreak $1 - \psi \geq \gamma$, or maximizing the outbreak size $1 - \psi$ for a given cost budget can be formulated. Theorem IV.1 can be easily extended to include the scenario $T_2 > T_1$, in this case we would obtain $\frac{d\psi}{dq} < 0$ and $\frac{d\bar{v}}{dq} > 0$. This result would allow the optimization problem to be solved by solving multiple linear programs.

V. CONCLUSION AND FUTURE WORK

In this article we studied the problem of containing rumors in a social network. More specifically, we considered a scenario where individuals are unable to distinguish between the rumor and the information message, and proposed a training mechanism to recruit and train individuals. By using a branching process approximation we calculated the size of the rumor outbreak and the conditions for the occurrence of such outbreaks. We then formulated an optimization problem for minimizing the expected recruitment and training cost while ensuring prevention of rumors. The optimization problem turned out to be nonlinear, and we showed that for a fixed quality of training, T_2 , the problem becomes a linear programming problem. This enabled us to solve the general problem

by solving a set of linear programs. The problem solution exhibited interesting properties, such as the sensitivity of low degree nodes to the intensity of rumor and the cost budget and the lack of sensitivity of high degree nodes to the same. The solution to the optimization problem exhibited a pattern which can be easily converted to an implementable heuristic. Furthermore, our results can be easily extended to address the information spreading problem.

More importantly, our results have implications on rumor control policies. Many Governments are concerned about dangerous rumor outbreaks and misinformation epidemics propagated on online social networks, and some, like the Indian Government and the Chinese Government have drafted policies [6, 25] for controlling them. However, these policies are drafted in an ad hoc manner. Furthermore they are not well studied and they do not work. In contrast, our approach gives guarantees on the outbreak size, uses the minimal possible resources and can be implemented in the form of a heuristic.

In this article we have assumed that only the rumor, modeled as a single message, is circulated in the social network. There may be situations where, messages other than the primary rumor message circulate in the social network. These messages may interact with the primary rumor message either helping or hindering its spread. Analysis of a model which incorporates such interactions may reveal novel approaches for combating rumor outbreaks. We leave this interesting problem to the future.

APPENDIX

Theorem IV.1 is established using Lemma A.2 and A.3, while Lemma A.1 is used in the proof of Lemma A.2 and A.3.

Lemma A.1. *For all $a, b \in [0, 1]$ and $k_1 + k_2 \leq n$, $n \in \mathbb{Z}^+$ and $f : \mathbb{Z} \rightarrow \mathbb{R}$ the following is true:*

$$\begin{aligned} & \sum_{k_1=0}^n \sum_{k_2=0}^{n-k_1} f(k_1 + k_2) k_2 \binom{k_1 + k_2}{k_2} a^{k_2-1} b^{k_1} \\ & - \sum_{k_1=0}^n \sum_{k_2=0}^{n-k_1} f(k_1 + k_2) k_1 \binom{k_1 + k_2}{k_2} a^{k_2} b^{k_1-1} = 0 \end{aligned}$$

Proof:

We can change switch the indices in the second term, i.e.,

$$\begin{aligned} & \sum_{k_1=0}^n \sum_{k_2=0}^{n-k_1} f(k_1 + k_2) k_1 \binom{k_1 + k_2}{k_2} a^{k_2} b^{k_1-1} \\ & = \sum_{k_1=0}^n \sum_{k_2=0}^{n-k_1} f(k_1 + k_2) k_2 \binom{k_1 + k_2}{k_2} a^{k_1} b^{k_2-1} \end{aligned}$$

Hence,

$$\begin{aligned} LHS &= \sum_{k_1=0}^n \sum_{k_2=0}^{n-k_1} f(k_1 + k_2) k_2 \binom{k_1 + k_2}{k_2} \left(a^{k_2-1} b^{k_1} - a^{k_1} b^{k_2-1} \right) \\ &= \sum_{k_1=0}^n \sum_{k_2=0}^{n-k_1} g(k_1, k_2) \end{aligned} \quad (3)$$

We now count the number of terms in the above equation and show that they are even. An expression indexed by a specific k_1 and k_2 denotes a term, e.g, $g(1, 1)$ is a term. The total number of terms in the summation $= \sum_{i=1}^{n+1} i = \frac{(n+1)(n+2)}{2}$. Out of those, $n+1$ terms are 0 due to the k_2 multiplier ($k_2 = 0$ for $k_1 = 0 \rightarrow n$). Additionally, when $k_2 = k_1 + 1$ equation (3) is zero. The total number of terms when $k_2 = k_1 + 1$ is given by $\lfloor \frac{n+1}{2} \rfloor$.

Since, these terms are zero, subtracting out these terms from the total number of terms results in

$$\begin{aligned} & \frac{(n+1)(n+2)}{2} - (n+1) - \left\lfloor \frac{n+1}{2} \right\rfloor \\ &= \frac{n^2}{2} \text{ for } n \text{ even} \\ &= \frac{(n-1)(n+1)}{2} \text{ for } n \text{ odd} \end{aligned}$$

Thus, the remaining terms are even for both n odd and even. This allows us to pair the terms. Consider one such pairing: the term with indices k_1, k_2 are paired with a term with indices \hat{k}_1, \hat{k}_2 where $\hat{k}_2 = k_1 + 1$ and $\hat{k}_1 = k_2 - 1$. If we sum these two terms we obtain

$$\begin{aligned} & g(k_1, k_2) + g(\hat{k}_1, \hat{k}_2) \\ &= f(k_1 + k_2) a^{k_2-1} b^{k_1} \left(\frac{k_2(k_1 + k_2)!}{k_1! k_2!} - \frac{k_2(k_1 + k_2)!}{k_1! k_2!} \right) \\ &+ f(k_1 + k_2) a^{k_1} b^{k_2-1} \left(\frac{(k_1 + 1)(k_1 + k_2)!}{(k_2 - 1)!(k_1 + 1)!} - \frac{(k_1 + 1)(k_1 + k_2)!}{(k_2 - 1)!(k_1 + 1)!} \right) \\ &= 0 \end{aligned}$$

Thus, the summation of the remaining terms is zero, which completes the proof. \blacksquare

Lemma A.2. *If $T_2 < T_1$ then \tilde{v} is strictly decreasing with respect to q , i.e, $\frac{d\tilde{v}}{dq} < 0$, $\forall q \in [0, 1]$.*

Proof:

$$\begin{aligned} & \frac{d}{dq} \tilde{v} = \\ & T_1 \sum_{k_1, k_2} k_1 Q(k_1 + k_2) \binom{k_1 + k_2}{k_2} \left(k_2 q^{k_2-1} r^{k_1} - k_1 q^{k_2} r^{k_1-1} \right) \\ & + T_2 \sum_{k_1, k_2} k_2 Q(k_1 + k_2) \binom{k_1 + k_2}{k_2} \left(k_2 q^{k_2-1} r^{k_1} - k_1 q^{k_2} r^{k_1-1} \right) \end{aligned}$$

where $r = 1 - q$. Let,

$$a_1 = \sum_{k_1, k_2}^{\infty} k_1 Q(k_1 + k_2) \binom{k_1 + k_2}{k_2} (k_2 q^{k_2-1} r^{k_1} - k_1 q^{k_2} r^{k_1-1})$$

$$a_2 = \sum_{k_1, k_2}^{\infty} k_2 Q(k_1 + k_2) \binom{k_1 + k_2}{k_2} (k_2 q^{k_2-1} r^{k_1} - k_1 q^{k_2} r^{k_1-1})$$

Adding a_1 and a_2 we get

$$a_1 + a_2 = \sum_{k_1, k_2}^{\infty} (k_1 + k_2) Q(k_1 + k_2) \binom{k_1 + k_2}{k_2} (k_2 q^{k_2-1} r^{k_1} - k_1 q^{k_2} r^{k_1-1})$$

Real world networks are always finite, hence $Q(k_1 + k_2) = 0$ for $k_1 + k_2 > k_{max}$, where k_{max} is the maximum degree. From Lemma A.1, $a_1 + a_2 = 0$. Now we prove that $a_2 > 0$. Let $k_1 + k_2 = m$.

$$a_2 = \sum_{m=1}^{k_{max}} Q(m) \left[\frac{1}{q} \sum_{k_2=0}^m k_2^2 \binom{m}{k_2} q^{k_2} r^{m-k_2} - \frac{1}{r} \sum_{k_1=0}^m k_1(m-k_1) \binom{m}{k_1} q^{m-k_1} r^{k_1} \right]$$

The summations are the second moments of a binomial random variable. $E[X^2] = Var[X] + E[X]^2$, $E[X] = mq$, $Var[X] = mqr$.

$$a_2 = \sum_{m=1}^{k_{max}} Q(m) \left[\frac{1}{q} (mqr + m^2 q^2) - \frac{1}{r} (m^2 r - mqr - m^2 r^2) \right]$$

$$= \sum_{m=1}^{k_{max}} Q(m) m$$

$$> 0$$

Since $T_2 < T_1$, $T_1 a_1 + T_2 a_2 < 0$, which completes the proof. ■

Lemma A.3. For $\psi \in (0, 1)$, if $T_2 < T_1$ then ψ is strictly increasing with respect to q , i.e., $\frac{d\psi}{dq} > 0$, $\forall q \in [0, 1]$.

Proof: Let, $\psi = g(u^*, q)$ where u^* is the solution of the fixed point equation $u = f(u, q)$.

$$g(u^*, q) = \sum_{k_1, k_2}^{\infty} \alpha^{k_1} \beta^{k_2} P(k_1 + k_2) \binom{k_1 + k_2}{k_2} q^{k_2} (1-q)^{k_1}$$

$$f(u, q) = \sum_{k_1, k_2}^{\infty} \alpha^{k_1} \beta^{k_2} Q(k_1 + k_2) \binom{k_1 + k_2}{k_2} q^{k_2} (1-q)^{k_1}$$

where $\alpha = 1 + (u^* - 1)T_1$ and $\beta = 1 + (u^* - 1)T_2$. We first show that the solution to the fixed point equation is strictly increasing with q . It can be easily shown that $\frac{\partial f(u, q)}{\partial u} > 0$ and $\frac{\partial^2 f(u, q)}{\partial u^2} > 0$ for all $u, q \in [0, 1]$. Thus f is a convex function in u for any fixed q . Also $f(0, q) > 0$ for all $q \in [0, 1]$.

$$\frac{\partial f(u, q)}{\partial q} =$$

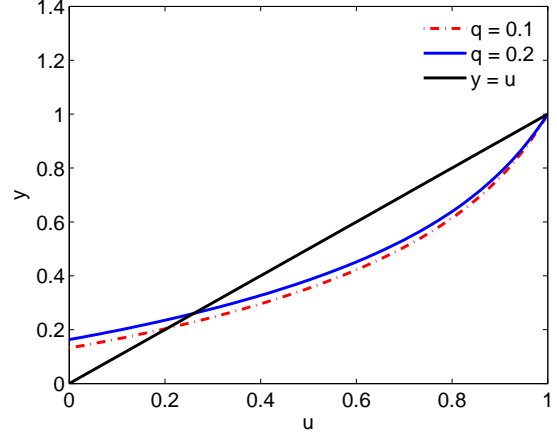


Fig. 8. Fixed point equation, $T_1 = 0.7$, $T_2 = 0.1$

$$\sum_{k_1, k_2}^{\infty} \alpha^{k_1} \beta^{k_2} Q(k_1 + k_2) \binom{k_1 + k_2}{k_2} (k_2 q^{k_2-1} r^{k_1} - k_1 q^{k_2} r^{k_1-1})$$

$$= \beta \sum_{k_1, k_2}^{\infty} Q(k_1 + k_2) \binom{k_1 + k_2}{k_2} k_2 (\beta q)^{k_2-1} (\alpha r)^{k_1}$$

$$- \alpha \sum_{k_1, k_2}^{\infty} Q(k_1 + k_2) \binom{k_1 + k_2}{k_2} k_1 (\beta q)^{k_2} (\alpha r)^{k_1-1}$$

From Lemma A.1,

$$\sum_{k_1, k_2}^{\infty} Q(k_1 + k_2) \binom{k_1 + k_2}{k_2} k_2 (\beta q)^{k_2-1} (\alpha r)^{k_1}$$

$$- \sum_{k_1, k_2}^{\infty} Q(k_1 + k_2) \binom{k_1 + k_2}{k_2} k_1 (\beta q)^{k_2} (\alpha r)^{k_1-1}$$

$$= 0$$

Now $\beta > \alpha$ because $T_2 < T_1$, which implies $\frac{\partial f(u, q)}{\partial q} > 0$.

Let u_0^* be the solution of the fixed point equation for some $q = q_0$ and let u_1^* be the solution to the fixed point equation when $q = q_1$ where $q_0 < q_1$. Also, u_1^* exists since we have assumed $\psi < 1$.

The curve $y = f(u, q_1)$ lies above the curve $y = f(u, q_0)$ and hence, $u_0^* < f(u_0^*, q_1)$, or in other words, the curve $y = f(u, q_1)$ has shifted above the original fixed point. Consider set I of points u such that $f(u, q_1) > f(u_0^*, q_1)$, clearly $u > u_0^*$ for all $u \in I$ (as $\frac{\partial f}{\partial u} > 0$). The line $y = u$, $\forall u \in I$ lies above the curve $y = f(u, q_0)$ for $u \geq u_0^*$ because $f(u)$ is convex in u and the point $u = 1$ is a fixed point (the line cannot be a tangent). Since, $f(u, q_1) > f(u, q_0)$, $\forall u \in [0, 1]$, there must exist a point u_1^* belonging to set I which lies on the line $y = u$, i.e., $u_1^* = f(u_1^*, q_2)$. Since $f(u, q)$ is continuous and differentiable in u, q $\frac{du^*}{dq} > 0$. This is illustrated in Fig. 8.

The function $g(u^*, q)$ has the same structure as the function $f(u, q)$, and hence using the same procedure it can be shown that $\frac{\partial g(u, q)}{\partial q} > 0$. The total derivative $\frac{d\psi}{dq}$ is given by:

$$\frac{d\psi}{dq} = \frac{\partial g}{\partial q} + \frac{\partial g}{\partial u} \frac{du^*}{dq}$$

Since all the terms on the right hand side of the above equation are positive, $\frac{d\psi}{dq} > 0$. ■

Theorem IV.1 follows from Lemma A.2 and A.3.

Lemma A.4. For a fixed T_1 and q , $\frac{d\psi}{dT_2} < 0$ for $\psi \in [0, 1)$.

Proof: Now, $\psi = g(u^*, T_2)$, u^* is the solution to the fixed point equation $u = f(u, T_2)$ where f and g are given by

$$g(u^*, T_2) = \sum_{k_1, k_2}^{\infty} \alpha^{k_1} \beta^{k_2} P(k_1 + k_2) \binom{k_1 + k_2}{k_2} q^{k_2} (1 - q)^{k_1}$$

$$f(u, T_2) = \sum_{k_1, k_2}^{\infty} \alpha^{k_1} \beta^{k_2} Q(k_1 + k_2) \binom{k_1 + k_2}{k_2} q^{k_2} (1 - q)^{k_1}$$

where $\alpha = 1 + (u^* - 1)T_1$ and $\beta = 1 + (u^* - 1)T_2$.

It is easy to show that $\frac{\partial f}{\partial T_2} < 0$ and $\frac{\partial g}{\partial T_2} < 0$. Using arguments similar to the ones described in Lemma A.3 one can write, $\frac{du^*}{dT_2} < 0$. The total derivative $\frac{d\psi}{dT_2}$ is given by:

$$\frac{d\psi}{dT_2} = \frac{\partial g}{\partial T_2} + \frac{\partial g}{\partial u} \frac{du^*}{dT_2}$$

Since $\frac{\partial g}{\partial T_2} < 0$, $\frac{du^*}{dT_2} < 0$ and $\frac{\partial g}{\partial u} > 0$, we have $\frac{d\psi}{dT_2} < 0$. ■

REFERENCES

- [1] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: understanding microblogging usage and communities," in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*. ACM, 2007, pp. 56–65.
- [2] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 497–506.
- [3] S. M. Friedman, "Three mile island, chernobyl, and fukushima: An analysis of traditional and new media coverage of nuclear accidents and radiation," *Bulletin of the Atomic Scientists*, vol. 67, no. 5, pp. 55–65, 2011.
- [4] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, "Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy," in *Proceedings of the 22nd international conference on World Wide Web companion*, 2013, pp. 729–736. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2487788.2488033>
- [5] C. A. Cassa, R. Chunara, K. Mandl, and J. S. Brownstein, "Twitter as a sentinel in emergency situations: lessons from the boston marathon explosions," *PLoS currents*, vol. 5, 2013.
- [6] U. Franke, "Disconnecting digital networks: A moral appraisal," *New ICTs and Social Media: Revolution, Counter-Revolution and Social Change*, vol. 18, p. 23, 2012.
- [7] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 137–146.
- [8] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '09, 2009, pp. 199–208. [Online]. Available: <http://doi.acm.org/10.1145/1557019.1557047>
- [9] W. Chen, Y. Yuan, and L. Zhang, "Scalable influence maximization in social networks under the linear threshold model," in *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 2010, pp. 88–97.
- [10] A. Karnik and P. Dayama, "Optimal control of information epidemics," in *Communication Systems and Networks (COMSNETS), 2012 Fourth International Conference on*. IEEE, 2012, pp. 1–7.
- [11] P. Dayama, A. Karnik, and Y. Narahari, "Optimal incentive timing strategies for product marketing on social networks," in *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*. International Foundation for Autonomous Agents and Multiagent Systems, 2012, pp. 703–710.
- [12] R. M. Tripathy, A. Bagchi, and S. Mehta, "A study of rumor control strategies on social networks," in *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010, pp. 1817–1820.
- [13] C. Budak, D. Agrawal, and A. El Abbadi, "Limiting the spread of misinformation in social networks," in *Proceedings of the 20th international conference on World wide web*. ACM, 2011, pp. 665–674.
- [14] N. P. Nguyen, G. Yan, M. T. Thai, and S. Eidenbenz, "Containment of misinformation spread in online social networks," in *Proceedings of the 3rd Annual ACM Web Science Conference*, ser. WebSci '12, 2012, pp. 213–222. [Online]. Available: <http://doi.acm.org/10.1145/2380718.2380746>
- [15] R. Cohen, S. Havlin, and D. Ben-Avraham, "Efficient immunization strategies for computer networks and populations," *Physical review letters*, vol. 91, no. 24, p. 247901, 2003.
- [16] J. Huang and X. Jin, "Preventing rumor spreading on small-world networks," *Journal of Systems Science and Complexity*, vol. 24, no. 3, pp. 449–456, 2011.
- [17] M. Khouzani, S. Sarkar, and E. Altman, "Dispatch then stop: Optimal dissemination of security patches in mobile wireless networks," in *Decision and Control (CDC), 2010 49th IEEE Conference on*. IEEE, 2010, pp. 2354–2359.
- [18] —, "Optimal propagation of security patches in mobile wireless networks," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 38, no. 1. ACM, 2010, pp. 355–356.
- [19] M. Molloy and B. Reed, "A critical point for random graphs with a given degree sequence," *Random Structures & Algorithms*, vol. 6, no. 2-3, pp. 161–180, 1995. [Online]. Available: <http://dx.doi.org/10.1002/rsa.3240060204>
- [20] T. E. Harris, *The Theory of Branching Processes*. Springer-Verlag, Berlin, 1963.

- [21] M. Newman, *Networks: An Introduction*. Oxford University Press, 2010.
- [22] R. Durrett, *Random graph dynamics*. Cambridge university press, 2007, vol. 20.
- [23] M. E. Newman, "Spread of epidemic disease on networks," *Physical review E*, vol. 66, no. 1, p. 016128, 2002.
- [24] G. Caldarelli, *Scale-free networks: complex webs in nature and technology*. Oxford University Press, 2007.
- [25] Reuters, "China threatens tough punishment for online rumor spreading." <http://www.reuters.com/article/2013/09/09/us-china-internet-idUSBRE9880CQ20130909>, Sep. 9, 2013.